Statistical Mechanical Formulation and Simulation of Prime Factorization of Integers

Chihiro H Nakajima*

[1]*Department of Physics, Kyushu University, 33 Fukuoka 812-8581, Japan*

We propose a new formulation of the problem of prime factorization of integers. With replica exchange Monte Carlo simulation, the behavior which is seemed to indicate exponential computational hardness is observed. But this formulation is expected to give a new insight into the computational complexity of this problem from a statistical mechanical point of view.

KEYWORDS: statistical mechanics, prime factorization, extended ensemble Monte Carlo method, computational complexity

1 Introduction

Prime factorization problem is one of relatively few problems called NP-intermediate, which is not considered to be NP-complete but no polynomial algorithm has been found. In this proceeding, we propose a statistical mechanical formulation of this problem and numerical analysis of them with Monte Carlo algorithm. The statistical mechanical modeling would reveal new features lying under the computational hardness of the problem through the structure of its phase space landscape [1–3]. In particular, one might determine the computational complexity class of the problem with an analysis of the existence and the order of phase transition phenomena of the model.

Furthermore, in terms of practicality, there have been several cases which the computational hardness of NP-hard problems are overcome by probabilistic algorithms. Therefore, also in practical point of view, it attracts our attention to studying the prime factorization problem with the form that is tractable by probabilistic algorithms which is now conventionally used in the field of statistical mechanics and verifying wheather it can be solved in polynomial time with them.

2 Models and Methods

2.1 Guidelines for Formulation

Suppose that an integer N_o is given. To obtain the prime factorization of N_o, we will solve an optimization problem by Markov chain Monte Carlo (MCMC) simulation with the cost function in the phase space. First, when an integer N_o is located in $2^{n-1} < N_o \leq 2^n$, the number of its prime divisors is bounded by n. In other words, a number n is defined for each N_o as

$$n = \lceil \log_2 N_o \rceil, \tag{2.1}$$

where $\lceil x \rceil$ is the minimal integer which is larger than x. Let $\{d_i\}$ be the state in the phase space composed of the set of integers d_i.

Second, the cost function should be designed to favor states in which the elements of $\{d_i\}$ are divisors of N_o and to take its lowest value when the entire set of $\{d_i\}$ is the prime factorization of N_o. Using the information of the residues we can design such cost function. By definition of the residue, each $\mod(N_o, d_i)$, which is obtained by dividing N_o by each d_i, takes the value 0 only in the case that d_i is a divisor of N_o, and becomes positive in the other case. Instead of $\mod(N_o, d_i)$ itself, we can also adopt following function of d_i and $\mod(N_o, d_i)$,

$$\varepsilon_i = \min\left(\ \mod(N_o, d_i),\ d_i - \mod(N_o, d_i)\ \right), \tag{2.2}$$

without loss of such property.

The order of the residue $\mod(N_o, d_i)$ or ε_i is up to that of $O(exp(n))$. To formulate a proper statistical mechanical model, we would like to keep the extensiveness of the cost function; Hamiltonian, with respect to n. By taking logarithm [see Eq.(2.5) below] or using coefficients of p-adic expansion ε_i [see Eq.(2.9)-(2.11) below], we can keep such extensiveness. Especially, with the case of using the p-adic expansion coefficient, the extensiveness is naturally guaranteed

---

and its value always becomes integer. Thus it is particularly convenient in order to calculate the statistical mechanical quantities.

2.2 Model Hamiltonian

The cost function is mainly composed of two contributions H_1 and H_2. H_1 comes from the residue terms and H_2 does from the difference of the product $\prod_i d_i$ from N_o. When representing the prime factorization form by the entire set of d_i we set $i = n$ and prepare n integers $\{d_i\} = \{d_1, \cdots d_n\}$. Each d_i takes the value $d_i \in \{1, \cdots, 2^n\}$.

Thus the detail of the resulting Hamiltonian $H_{whole}(\{d_i\})$ is shown as,

$$H_{whole}(\{d_i\} \mid N_o) = H_1 + H_2 - \gamma M, \qquad (\gamma \geq 0) \tag{2.3}$$

$$H_1 = \sum_{i=1}^{n} \log(1 + \varepsilon_i), \tag{2.4}$$

$$H_2 = \frac{1}{n^2}\left(\log N_o - \sum_{i=1}^{n} \log(d_i)\right)^2, \tag{2.5}$$

where M is the number of integers included in $\{d_i\}$ which is larger than 1. H_1 takes the value 0 when all d_i become any divisors of N_o. This term works to each d_i locally. On the other hand, H_2 takes the value 0 when the product of all d_i is equal to N_o. This term works globally to prohibit the case that all d_i takes the value 1. Thus the Hamiltonian which takes the value 0 if and only if the full set of prime divisors is realized by $\{d_i\}$.
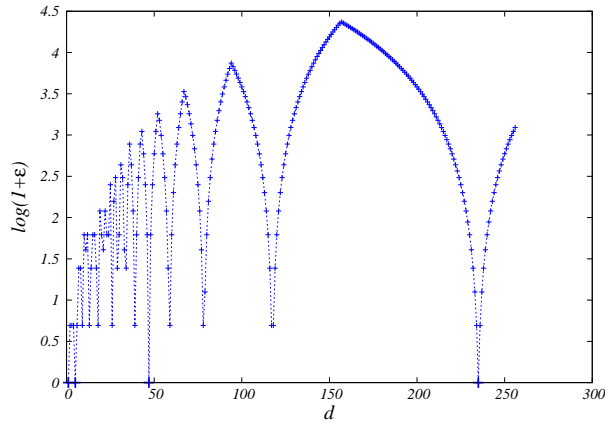

Fig. 1: The profile of $\log(1 + \varepsilon_i)$ introduced in Eq.(2.4) in the case of $N_o = 235$. It indicates that Eq.(2.4) has several local minima.
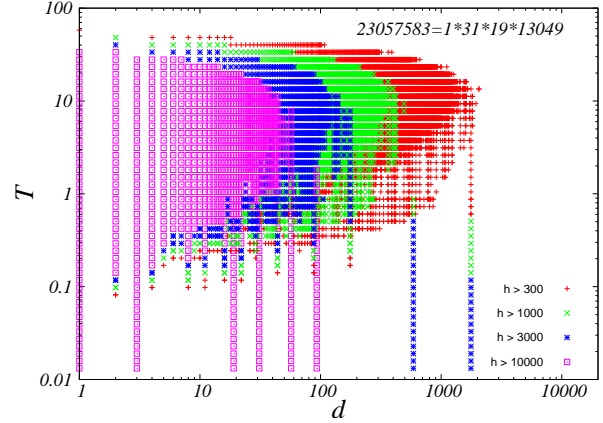

Fig. 2: The histogram of the result of the simulation of H_{whole} with replica exchange Monte Carlo method in the case of $N_o = 23057583$. The value of γ is taken to be 0 in this result. The color red, green, blue, and purple represent the nmber ogf hit more than $300, 1000, 3000$ and 10000 times respectively.At sufficiently low temperature, only the case which d_i is a divisor of N_o is sampled.

2.3 Another model

To formulate the problem of prime factorization as ground state searching, there is some arbitrariness in designing of Hamiltonian. In practice, it is also efficient to decompose N_o into two divisors recursively and apply the primality test. In this procedure, the Hamiltonian of factorization of N_o into d_1, d_2 can be written as follows,

$$H_{elem}(\{d_i\} \mid N_o) = H_1 + H_2, \tag{2.6}$$

$$H_1 = \sum_{j=1}^{n}\left(\sigma(\varepsilon_1, j) + \sigma(\varepsilon_2, j)\right), \tag{2.7}$$

$$H_2 = \sum_{k=1}^{n} \sigma(|d_1 d_2 - N_o|, k), \tag{2.8}$$

where the function σ represents the coefficient of the p-adic expansions of the variables, defined as,

$$\varepsilon_i = \sum_j \sigma(\varepsilon_i, j)p^{j-1}, \tag{2.9}$$

$$|d_1 d_2 - N_o| = \sum_k \sigma(|d_1 d_2 - N_o|, k)p^{k-1}, \tag{2.10}$$

$$\sigma \in \{0, \cdots, p-1\}. \tag{2.11}$$

The ground states of this Hamiltonian does not become the prime factorization of N_o unless it is the composite number of two prime numbers. But we can investigate the elemental process of the factorization with this Hamiltonian.

2.4 Searching with Replica exchange Monte Carlo Method

Changing each d_i randomly with each step, we search the prime factorization of N_o by optimizing the state $H_{whole}(\{d_i\}|N_o)$ or $H_{elem}(\{d_i\}|N_o)$ with certain condition. We have implemented the rule of the transitions in the phase space by representing each d_i with p-adic expansion similar to Eq.(2.9)-(2.11) in the cases of the cost function. But in this case we adopted following form of the expansion,

$$d_i = 1 + \sum_j \sigma(d_i - 1, j)q^{j-1}, \tag{2.12}$$

$$\sigma \in \{0, \cdots, q-1\}, \tag{2.13}$$

so that each d_i does not take the value 0. We note q instead of p to avoid the confusion. Thus the phase space is divided-and-conquered by Potts (Ising) like variables σ. Transitions in the phase space are performed by shifting (or flipping) the value of each σ. For example, when we adopt $q = 2$, the above two Hamiltonians H_{whole} and H_{elem} are described with n^2 and $2n$ degrees of freedom respectively.

In the potential energy landscape of H_{whole} or H_{elem}, like that of spin glass models, there are several local minima [see Fig.1]. Even with probabilistic sampling, in ordinary way, the random walker can easily become trapped in each minima. To avoid these trap, it requires heating and annealing of the system. Thus we perform simulations with several temperature in parallel with exchanging [5] the correspondence of each walker and temperature. When a replica is in temperature T_1, the state is sampled as an instance in distribution $P(\Gamma, T_1)$. With some certain condition of transition probability, we can keep the canonical stationary distribution in each temperature.

Replica exchange (or some other extended ensemble) Monte Carlo methods are applied to several optimization or constraint satisfaction problems branching from spin glass, including NP-hard problems, and powerful tool for estimating expectation values with less systematic errors, finding the optimal solution, computing entropy or free energy, counting the number of solutions of these models.

3 Result

3.1 Behavior of Computational Cost

We numerically observed the first passage time τ_{first}, the Monte Carlo step that the walker of the MCMC simulation visits the state of the correct factorization for the first time, and its dependence on the system size n both for H_{whole} and H_{elem}. Various samples of N_o are generated by multiplying two prime numbers which are randomly choosed but moderately close to each other.

First we start from the explanation of the results of the simulations of H_{whole}. Fig.3 shows the dependence of τ_{first} on $n_o = \log_2 N_o$. Here the results are obtained with $\gamma = 0$. The green points shows the average over 10 independent samples for each N_o. The average is taken as $\exp\left(\overline{\log \tau_{first}}\right)$ and the dispersion $\exp\left(\sqrt{\overline{\log \tau_{first}}^2 - \left(\overline{\log \tau_{first}}\right)^2}\right)$ indicates a measure of the variation. Due to the large variation, it is difficult to distinguish between the exponential dependence and power-law dependence with large exponent from these results. But even if the power-law dependence was correct, the exponent is estimated as nearly 8. Such a large value is thought to be quantitatively unfamiliar.

Fig.4 and 5 show the results of the case of H_{elem}. In this case we performed with two different manner of conquering the phase space, $q = 2$ (by Ising variables) and $q = 3$ (by Potts variables), with same N_o. The number of degrees of freedom and the transitions in the phase space are different between these cases. In both cases the results are obtained by averaging over 10 sets of independent simulations for each N_o. Even though the detailed structure of the phase space landscape is different, these results exhibit similar behavior to the case of H_{whole}. They are also seen to exhibit exponential dependence on n_o and large (still larger) variation in each sample.
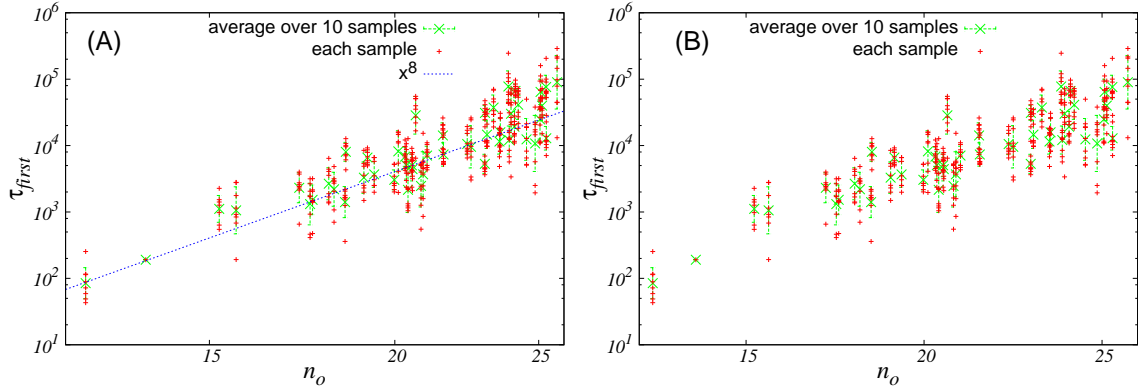
Fig. 3: The dependecce of τ_{first} on $n_o = \log_2 N_o$ with the simulation of H_{whole}. (A):Double-logarithmic plot, (B):Semilogrithmic plot. The red points replesents independent 10 samples for each $n_o = \log_2 N_o$ and the green points represents average over them. The error bar represents the dispersion obtained from $\log \tau_{first}$ of each simulation. It indicates a measure of variation.
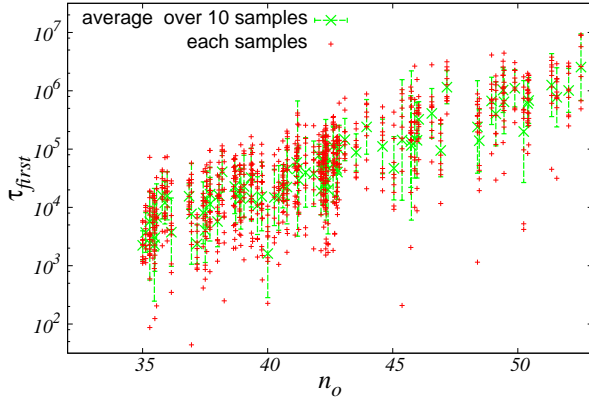

Fig. 4: The dependecce of τ_{first} on n_o with the simulation of H_{elem} with $q = 2$. The means of red points, green points and error bars are the same as Fig.3 respectively. Both the average and the dispersion are calculated with the same manner.

Fig. 5: The dependecce of τ_{first} on n_o with the simulation of H_{elem} with $q = 3$. The means of red points, green points and error bars are the same as Fig.3 respectively. Both the average and the dispersion are calculated with the same manner.
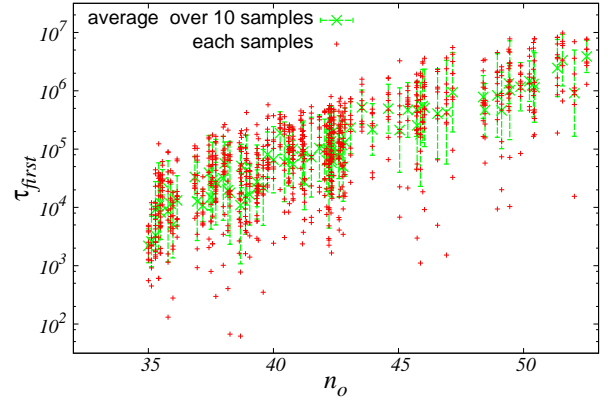
4   Summary and Discussion

In this proceeding we have proposed statistical mechanical formulations of the problem of factorization of integers. We set up two kinds of Hamiltonians and gave rough overviews of the size dependence of their computational cost for optimization by replica exchange Monte Carlo method. Though the results are not yet sufficient to obtain quantitative conclusion, we observed the behavior which is seemed to indicate that they require exponential computational cost. However, it should be noted that several questions and issues still remain.

First, the roughly yielded results of the first passage time should be improved by extensive investigation. An accurate determination of the probability distribution of the first passage times is left for future work. As the system size dependence of the distribution directly reflects the computational complexity of this problem, it should be precisely computed. In the above result about τ_{first}, the dispersion of $\log \tau_{first}$ calculated from 10 independent samples is given as a measure of the spread of the distribution function, assuming that their distribution is a Gaussian. But the tendency of the variation in the figures indecates a possibility that the actual form of the distribution may not be Gaussian.

Second, it should be emphasized that the number field sieve method is already known as an relatively efficient algorithm that achieves $O(\exp(n^{\frac{1}{2}}))$ computational cost. This fact is one of the reasons for that the prime factorization problem is expected to be not NP problem. Considering the fact, it is suggested that the formulation in this proceeding has not yet reached the level of the maximum possible efficiency in the classical computer. In order to use the above results as any evidence about the computational complexity class with classical probabilistic algorithm, it is thought to be the most reasonable with the case that is confirmed with the most efficient algorithm. We argue that it is still not excluded out the possibility that we can reach the most efficient level of computation amount by the choice of the way of dividing-and-conquering the phase space, the transition rule, and the temperature points of each replicas [7]. While, it is non-trivial whether we can introduce Markov chain and cost function into the algorithm of the number field sieve

method.

As an another subject of future works, we plan to investigate the static quantities of these statistical mechanical models. In the statistical mechanics of spin glass theory, the relationship among the phase transition, computational complexity, and the structure of the potential energy landscape are discussed [1,3,4]. In this direction, a number of studies have been investigated mainly around the adaptation of the cavity method to random satisfiability, random graph coloring and several other NP-hard problems over this ten and a few years [2,4,6]. In the simulations with above Hamiltonians, it seemed that the random walker in the phase space had been trapped in several isolated local minima which take the value 1 or 2 even in the case that N_o is composed of two prime numbers, when the number of true ground states are order of $O(1)$ in the above Hamiltonians. This property is expected to be related to the rapidly growing computational effort even with probabilistic algorithm. It would be effective to observe the density of states focusing on the asymptotic n dependence of the low energy tail. The characterics of the density of states can also appear in the temperature dependence of specific heat and entropy. By the detailed analysis of the above quantities, one could gain an understanding of the computational complexity of this problem.

REFERENCES

[1] R. Monasson, R. Zecchina, S. Kirkpatrick, B. Selman, and L. Troyansky "Determining computational complexity from characteristic 'phase transitions'," *Nature*, **400**: 133-137 (1999).

[2] M. Mezard, G. Parisi, and R. Zecchina "Analytic and algorithmic solution of random satisfiability problems," *Science*, **297**: 812-815 (2002).

[3] O.C.Martin, R. Monasson, and R. Zecchina "Statistical mechanics methods and phase transitions in optimization problems," *Theor. Comput. Sci.*, **265**: 3-67 (2001).

[4] M. Mezard and A. Montanari, *Information, Physics, and Computation*, Oxford University Press, Oxford, U.K. (2009).

[5] K. Hukushima and K. Nemoto, "Exchange Monte Carlo Method and Application to Spin Glass Simulations," *J. Phys. Soc. Jpn.*, **65**: 1604-1608 (1996).

[6] L. Zdeborova and F. Krzakala, "Phase transitions and computational difficulty in random constraint satisfaction problems," *J. Phys.: Conf. Ser.*, **95**: 012012 (2008).

[7] K. Klemm, A. Mehta, P. F. Stadler "Landscape encodings enhance optimization," *PLoS ONE*, **7(4)**: e34780 (2012).